THU-SSS-CCT-04

LEVERAGING PUBLIC MEDIA FOR LOW-RESOURCE TEXT-TO-SPEECH: A CASE STUDY OF VIETNAMESE LANGUAGE [5]

Nhi Pham, Erasmus Student
Department of Computer Systems and Technologies
Faculty of Electrical Engineering, Electronics and Automation
University of Ruse "Angel Kanchev"
E-mail: nhi.pt205190@sis.hust.edu.vn

Dr. Nguyen Thi Thu Trang, PhD
Department of Computer Science
School of Information and Communications Technology
Hanoi University of Science and Technology
E-mail: trangntt@soict.hust.edu.vn

Assoc. Prof. Adriana Borodzhieva, PhD
Department of Telecommunications,
Faculty of Electrical Engineering, Electronics and Automation
University of Ruse "Angel Kanchev"
Tel.: +359 82 888 734
E-mail: aborodzhieva@uni-ruse.bg

Abstract: *The performance gap between high-resource and low-resource Text-to-Speech (TTS) systems remains a significant challenge in speech synthesis. While mainstream models achieve high naturalness and intelligibility, they rely on massive amounts of high-fidelity, studio-recorded data, a requirement that is often impractical for many low-resource languages. In this work, we introduce an end-to-end data construction pipeline that leverages publicly available media for speech data collection. Using Vietnamese as a case study, we combine this data construction framework with a multilingual transfer learning approach to optimize model training efficacy. Experimental results indicate that integrating diverse, naturalistic public media significantly improves prosodic naturalness and speaker diversity compared to models trained on limited curated datasets. Our findings demonstrate a scalable, cost-effective approach for developing high-quality TTS in low-resource linguistic contexts without the need for specialized recording sessions.*

Keywords: *Speech Synthesis, Text-to-Speech, Low-Resource Language, Data Construction Pipeline, Transfer Learning.*

1. INTRODUCTION

Text-to-Speech (TTS) has seen significant advancements over the past decade, transitioning from concatenative methods to end-to-end neural architectures capable of generating nearly indistinguishable from human recordings in terms of both intelligibility and naturalness. However, the success of these deep learning-based systems heavily depends on the size and quality of the data. While high-resource languages benefit from extensive corpora such as LibriTTS (Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. Y., Jia, Y., Chen, Z., and Wu, Y., 2019) and AISHELL-3 (Shi, Y., Bu, H., Xu, X., Zhang, S. & Li, M., 2020), low-resource languages face a significant bottleneck. The scarcity of curated speech data not only limits model convergence but often results in synthesized speech that lacks prosodic variation and speaker diversity.

[5] Докладът е представен на студентската научна сесия на 08.05.2025 г. в секция „Комуникационна и компютърна техника" с оригинално заглавие на английски език: Leveraging Public Media for Low-Resource Text-To-Speech: A Case Study of Vietnamese Language.

This disparity is particularly evident in Vietnamese, a tonal language where lexical meaning is intrinsically linked to pitch contours. Existing publicly available Vietnamese corpora, such as VIVOS (Azzarelli, A., Gao, G., Kwan, H. M., Zhang, F., Anantrasirichai, N., Moolan-Feroze, O., and Bull, D., 2025), provide approximately 15 hours of speech across 46 speakers. However, this data is predominantly composed of audiobook readings characterized by monotonic prosody and controlled acoustic environments. Consequently, models trained on such restricted domains struggle to generalize to spontaneous, in-the-wild speech patterns, limiting their applicability in real-world scenarios. Collecting studio-quality data for low-resource languages is often expensive and logistically complex, necessitating alternative data acquisition strategies.

To bridge this gap, this work explores the viability of leveraging public media as a primary source for low-resource TTS development. Public broadcasting and user-generated content offer a vast repository of diverse speakers and naturalistic prosody, yet they introduce challenges regarding noise and transcription alignment. Therefore, we propose an adaptable, end-to-end dataset construction pipeline designed to process in-the-wild recordings effectively. This pipeline integrates automated noise suppression, strict speaker verification, and transcription alignment to ensure data quality comparable to curated corpora.

Using Vietnamese as a case study, we construct a new dataset comprising 51 hours of speech from 299 unique speakers sourced from public media. Furthermore, we implement a multilingual transfer learning approach to maximize the utility of this crawled data during model training. Experimental results demonstrate that integrating diverse, naturalistic public media significantly enhances prosodic naturalness and speaker diversity compared to models trained solely on limited curated datasets.

The remainder of this paper is organized as follows: Section 2 details the proposed data construction pipeline, including data crawling, cleaning, and labelling processes. Section 3 presents the experimental setup and evaluates the performance of TTS models trained on the new dataset. Finally, Section 4 concludes the paper with a summary of findings and future directions.

2. DATA CONSTRUCTION PIPELINE

2.1. Overall Pipeline

To address the scarcity of diverse speech data for Vietnamese TTS, we propose an automated, end-to-end data construction pipeline. This framework transforms raw public media into a structured speech-text corpus suitable for training multi-speaker TTS models. Fig. 1 illustrates the overall architecture of the pipeline, which consists of three primary stages: data acquisition, preprocessing and cleaning, and speaker labelling with text normalization.

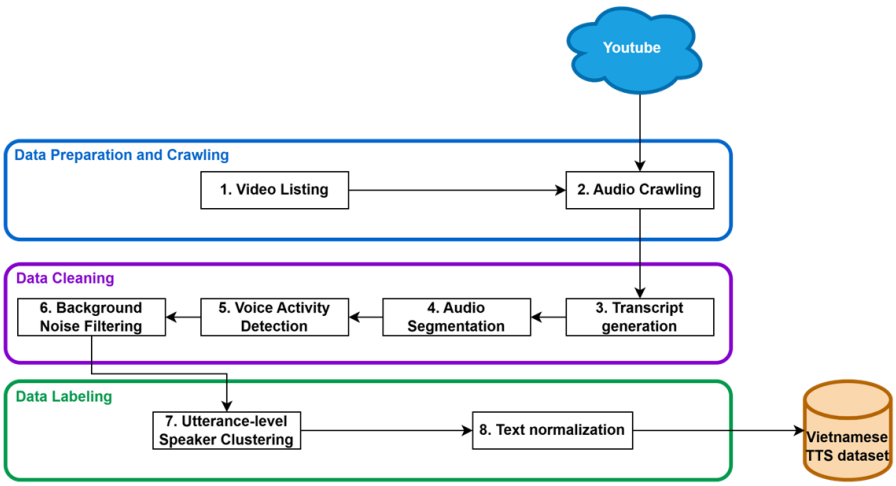

Fig. 1. The overall data construction pipeline for the Vietnamese dataset

 Copyrights© 2025 ISBN 978-954-712-978-8 (Print); e-ISBN 978-954-712-979-5

2.2. Data Preparation and Crawling

The data collection process employs a semi-automated strategy to ensure content relevance while maximizing scale. We initially curated a list of Vietnamese YouTube channels focusing on podcasts and audiobooks. Using this list, we automatically crawled public playlists containing keywords such as "audiobook" or "podcast" in the title. A key assumption in our crawling strategy is that individual playlists typically feature a single predominant speaker. This allows us to group crawled videos by playlist, facilitating initial speaker separation before fine-grained verification. This approach yielded a raw corpus significantly larger than existing curated datasets, providing a broad foundation for subsequent filtering.

2.3. Data Preprocessing and Cleaning

Raw media files often contain noise, music, and non-speech segments that degrade TTS training quality. We implemented a cleaning stage to ensure high-fidelity audio-text pairs. First, we utilized a pretrained Vietnamese Automatic Speech Recognition (ASR) model (Le, T.-T., Nguyen, L. T., and Nguyen, D. Q., 2024) to generate transcripts and extract word-level timestamps for each video. This forced alignment allows us to segment continuous audio into utterance-level clips corresponding to specific text spans. To minimize noise and misalignment, we applied several filtering criteria. Transcripts containing music markers, non-ASCII characters, or URLs were discarded to avoid non-speech artifacts. Additionally, text chunks corresponding to annotations, such as laugh, music, or applause, were removed from the transcript and aligned audio segments. Finally, segments shorter than five seconds were excluded. This threshold ensures that each utterance contains sufficient prosodic context for effective TTS training while minimizing truncation artifacts.

2.4. Speaker Labelling and Clustering

Accurate speaker identification is critical for multi-speaker TTS models. Since public media lacks consistent speaker IDs, we employed an unsupervised clustering approach followed by verification. For each video, we sampled five audio clips to calculate an average speaker embedding. We utilized the pretrained ECAPA-TDNN speaker encoder (Desplanques, B., Thienpondt, J., and Demuynck, K., 2020), known for its robustness in variable acoustic conditions, to extract these embeddings. Subsequently, we employed HDBSCAN (McInnes, L., Healy, J., and Astels, S., 2017), a density-based clustering algorithm, to group video-level embeddings into distinct speaker clusters. This method allows for the identification of speakers without pre-defining the number of clusters.

To validate the clustering accuracy, we constructed a verification set of 5 000 randomly sampled audio pairs assigned pseudo-labels based on the clustering results. Half of these pairs were manually inspected for speaker consistency, while the remaining pairs were evaluated using a pretrained ECAPA-TDNN speaker verification model fine-tuned on the Vietnam-Celeb dataset (Desplanques, B., Thienpondt, J., and Demuynck, K., 2020). Incorrect pairings identified through either method were corrected, and outlier samples were removed from their respective clusters. This process resulted in a final corpus of 299 distinct speakers.

2.5. Text Normalization

To ensure consistency between the input text and the acoustic model's expectations, we applied standard text normalization procedures tailored for Vietnamese. All text was converted to lowercase, and punctuation marks were removed to simplify the input vocabulary. Common abbreviations were expanded using a custom lookup list, such as converting "TP." to "thành phố". This list was initially generated using large language models and subsequently verified by native speakers. Furthermore, numeric digits found in captions, such as "18 giờ 5 phút", were converted to their spoken Vietnamese forms, such as "mười tám giờ năm phút", to ensure correct pronunciation during synthesis.

3. EXPERIMENTS AND RESULTS

3.1. Dataset

Following the construction pipeline described in Section 2, we obtained a comprehensive Vietnamese speech dataset comprising 299 distinct speakers and over 11 000 utterances. The total duration of the curated corpus is 51 hours, providing a substantial increase in volume compared to existing low-resource benchmarks. All audio utterances were resampled to 48 000 Hz to ensure high-fidelity synthesis compatibility. This dataset serves as the primary training resource for our proposed low-resource TTS framework, offering greater speaker diversity and prosodic variation than previously available corpora.

3.2. Model Architecture and Experimental Setup

For our experiments, we adopted the multilingual TTS architecture provided by IMS Toucan (Lux, F., Koch, J., Meyer, S., Bott, T., Schauffler, N., Denisov, P., Schweitzer, A., and Vu, N. T., 2023). Training a model from scratch for a new language is often time-consuming and computationally expensive; therefore, we leveraged a pretrained English model and adapted it to the Vietnamese language using our newly constructed dataset via transfer learning. Fig. 2 illustrates the architecture of the IMS Toucan (Lux, F., Koch, J., Meyer, S., Bott, T., Schauffler, N., Denisov, P., Schweitzer, A., and Vu, N. T., 2023) model used in this study. Training was conducted for up to 500 000 steps on a single NVIDIA Tesla V100 GPU with 32 GB of memory, utilizing a batch size of 32. The training process began with a warm-up phase of 200 steps, followed by a cosine schedule decay starting from an initial learning rate of 1e-4. To synthesize waveforms from the generated mel spectrograms, a HiFi-GAN vocoder (Kong, J., Kim, J., and Bae, J., 2020) was employed. To further enhance speaker identity in the synthetic speech compared to the reference, a fine-tuning phase of 150,000 steps was conducted using the monolingual dataset for the target speaker. During this fine-tuning phase, the learning rate was reduced to 6e-8, and the style encoder was frozen to preserve linguistic content while adapting speaker characteristics.

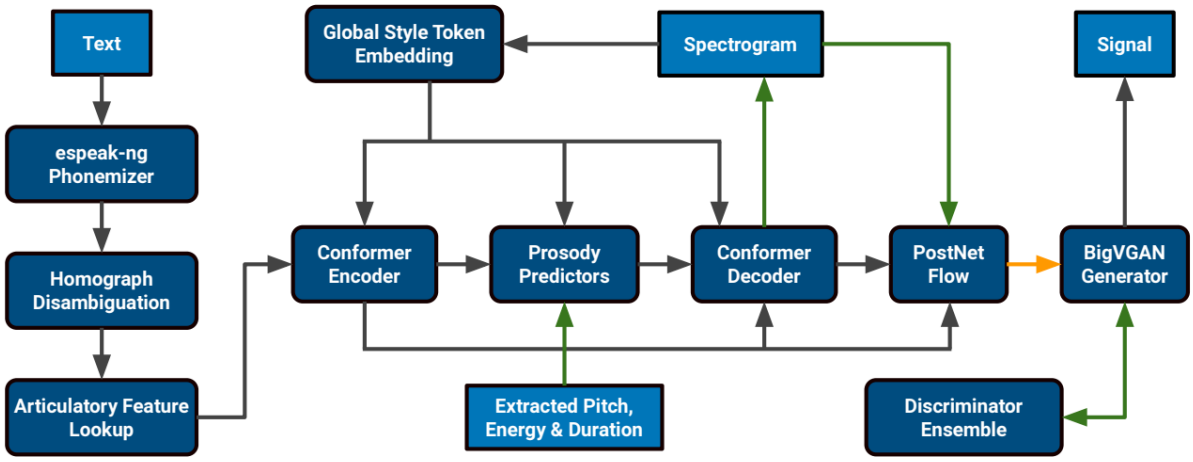

Fig. 2. Overview of all the components in the system: the green arrows show the losses applied at training time, the orange arrow only exists during inference, and the gradient is not passed through at training time (Lux, F., Koch, J., Meyer, S., Bott, T., Schauffler, N., Denisov, P., Schweitzer, A., and Vu, N. T., 2023)

3.3. Evaluation Metrics

We employed both subjective and objective evaluation metrics to assess the quality of the synthesized speech. For subjective evaluation, we utilized the Mean Opinion Score (MOS) to

measure the naturalness of the synthetic speech. Participants rated the speech samples on a scale from 1, indicating bad or very unnatural quality, to 5, indicating excellent or very natural quality.

MOS is a numerical measure of perceived quality, most commonly used for: 1) speech quality (telecom, VoIP); 2) audio quality (codecs, streaming); 3) video quality (streaming, compression). It reflects human subjective judgment, not machine-measured metrics. MOS is simply the average of all user ratings. If listeners/viewers rate quality on a 1 – 5 scale:

$$MOS = \frac{\sum_{i=1}^{N} R_i}{N}, \tag{1}$$

where R_i is the rating from person *i*, and *N* is the number of raters.

Additionally, the Similarity Mean Opinion Score (SMOS) was used to evaluate the similarity of the synthetic speech to a reference speaker, where participants rated the speech's resemblance to the target speaker's voice.

Similarity Mean Opinion Score (SMOS) is a subjective evaluation metric used mainly in speech synthesis and voice conversion. Instead of rating quality, listeners rate how similar the generated voice is to a target speaker. It answers the question: "How close does this sound to the original speaker?" The calculation is extremely simple – it is just the **average of listener similarity ratings:**

$$SMOS = \frac{\sum_{i=1}^{N} S_i}{N}, \tag{2}$$

where S_i is the similarity rating from listener *i*, and *N* is the number of listeners.

The evaluation involved 10 participants aged between 22 and 35 years old. All evaluations were conducted in a quiet environment to ensure optimal listening conditions and minimize external noise interference.

For objective evaluation, we implemented speaker verification and Word Error Rate (WER) metrics to quantify the acoustic fidelity and linguistic intelligibility of the synthesized audio, respectively.

Word Error Rate (WER) is the standard metric for evaluating speech-to-text accuracy. WER measures how different an automatic transcription is from the correct (reference) text. It is based on the minimum number of edits needed to transform the system output into the reference. Those edits are:

- *S* = Substitutions (wrong word instead of the correct one)
- *D* = Deletions (a word missing)
- *I* = Insertions (an extra word added)

Then WER is calculated as follows:

$$WER = \frac{S + D + I}{N}, \tag{3}$$

where: *S* is the number of substitutions, *D* is the number of deletions, *I* is the number of insertions, and *N* is the total number of words in the reference text. WER is usually expressed as a percentage.

To measure speaker similarity, we employed a pretrained ECAPA-TDNN model (Desplanques, B., Thienpondt, J., and Demuynck, K., 2020) trained on 187 hours of Vietnamese speech from 835 speakers in the Vietnam-Celeb dataset to extract x-vectors. An averaged x-vector of all utterances for each speaker in the constructed Vietnamese dataset and the

corresponding synthetic audio were extracted to calculate the cosine distance. A higher cosine similarity indicates a more similar speaker timbre between the synthesized and reference speech. For intelligibility assessment, a pretrained Whisper Large V3 (Le, T.-T., Nguyen, L. T., and Nguyen, D. Q., 2024) model was used to transcribe the synthetic speech into hypothesis texts for WER calculation.

To evaluate how closely the synthesized voice matches the vocal characteristics of the target speaker, we first measure speaker similarity using x-vector embeddings extracted with a pretrained ECAPA-TDNN model. When x-vectors are extracted from the *reference* speaker audio and the *synthetic* (generated) audio, two fixed-length speaker embeddings are obtained: x_{ref}, x_{syn}. To measure how similar the two voices are, the cosine similarity is computed:

$$Sim_{\cos} = \frac{x_{ref} \cdot x_{syn}}{\|x_{ref}\| . \|x_{syn}\|} \quad (4)$$

The numerator of (4) is the dot product of the two embeddings. The denominator is the product of their magnitudes. The result ranges from – 1 to 1. Higher values (closer to 1) mean the synthetic voice has a more similar timbre to the reference speaker.

3.4. Results and Discussion

The experimental results detailed in Table 1 demonstrate the effectiveness of the proposed dataset construction pipeline for Vietnamese Text-to-Speech (TTS) synthesis. The system achieved a Mean Opinion Score (MOS) of 3.24, reaching 96 % of the ground truth naturalness for Vietnamese text scenarios. For speaker similarity, the model recorded an objective score of 0.79 and a Similarity MOS (SMOS) of 3.15, which represents 84 % of the performance of the ground truth recordings. These metrics are particularly significant given that the dataset was developed from diverse public media sources rather than controlled studio environments. Furthermore, the system yielded a low Word Error Rate (WER) of 1.90 %, indicating a high level of linguistic accuracy and intelligibility derived from the automated labelling process.

Table 1. Evaluation results on the Vietnamese dataset with the proposed construction pipeline

Metrics / Dataset	***MOS***	***SMOS***	***Sim$_{cos}$***	***WER***
Ground truth	4.02	3.71	0.87	-
Vietnamese dataset (with the proposed pipeline)	3.24	3.15	0.79	1.90

CONCLUSION

This paper presents an automatic framework for developing high-quality Text-to-Speech (TTS) systems in low-resource linguistic environments by leveraging the availability of public media. By addressing the limitations of existing datasets, which are often restricted in scale and prosodic variety, we successfully constructed a 51-hour Vietnamese corpus featuring 299 diverse speakers sourced from "in-the-wild" YouTube content. Our three-phase construction pipeline, incorporating automated data crawling, data cleaning, and data labelling, effectively bridges the resource gap for the Vietnamese language without the need for expensive, specialized recording sessions.

The experimental results highlight the success of the proposed framework, as the model trained on the generated dataset achieved a Mean Opinion Score (MOS) of 3.24 for naturalness, representing 96 % of the ground truth performance for Vietnamese text scenarios. In terms of

speaker similarity, the model reached 84 % of the ground truth level, supported by an objective similarity score of 0.79 and a Similarity MOS (SMOS) of 3.15. Furthermore, a low Word Error Rate (WER) of 1.90 % confirms the high intelligibility and linguistic accuracy of the synthesized speech.

In the future, we aim to extend the dataset by incorporating additional audio to achieve a more comprehensive dialect balance within the Vietnamese language. We also plan to introduce more advanced processing modules to further refine data quality and apply this pipeline to other under-resourced languages to evaluate its robustness across diverse linguistic contexts.

REFERENCES

Azzarelli, A., Gao, G., Kwan, H. M., Zhang, F., Anantrasirichai, N., Moolan-Feroze, O., and Bull, D. (2025). "ViVo: A Dataset for Volumetric Video Reconstruction and Compression". arXiv preprint, arXiv:2506.00558.

Desplanques, B., Thienpondt, J., and Demuynck, K. (2020). "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification". arXiv preprint, arXiv:2005.07143.

Kong, J., Kim, J., and Bae, J. (2020). "HiFi-GAN: Generative Adversarial Networks for Efficient and High-Fidelity Speech Synthesis". Advances in Neural Information Processing Systems, vol. 33, pp. 17022-17033, https://arxiv.org/abs/2010.05646.

Le, T.-T., Nguyen, L. T., and Nguyen, D. Q. (2024). "Phowhisper: Automatic Speech Recognition for Vietnamese". arXiv preprint, arXiv:2406.02555.

Lux, F., Koch, J., Meyer, S., Bott, T., Schauffler, N., Denisov, P., Schweitzer, A., and Vu, N. T. (2023). "The IMS Toucan System for the Blizzard Challenge", arXiv preprint, arXiv:2310.17499.

McInnes, L., Healy, J., and Astels, S. (2017). "hdbscan: Hierarchical Density-Based Clustering", Journal of Open Source Software, vol. 2 (11), pp. 205, doi:10.21105/joss.00205.

Shi, Y., Bu, H., Xu, X., Zhang, S. & Li, M. (2020). "AISHELL-3: A Multi-speaker Mandarin TTS Corpus and the Baselines", doi: 10.48550/arXiv.2010.11567.

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. Y., Jia, Y., Chen, Z., and Wu, Y. (2019). "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech". arXiv preprint, arXiv:1904.02882.